

729G78 Artificiell intelligens

Maskininlärning - Beslutsträd

Robin Keskisärkkä

HCS/IDA

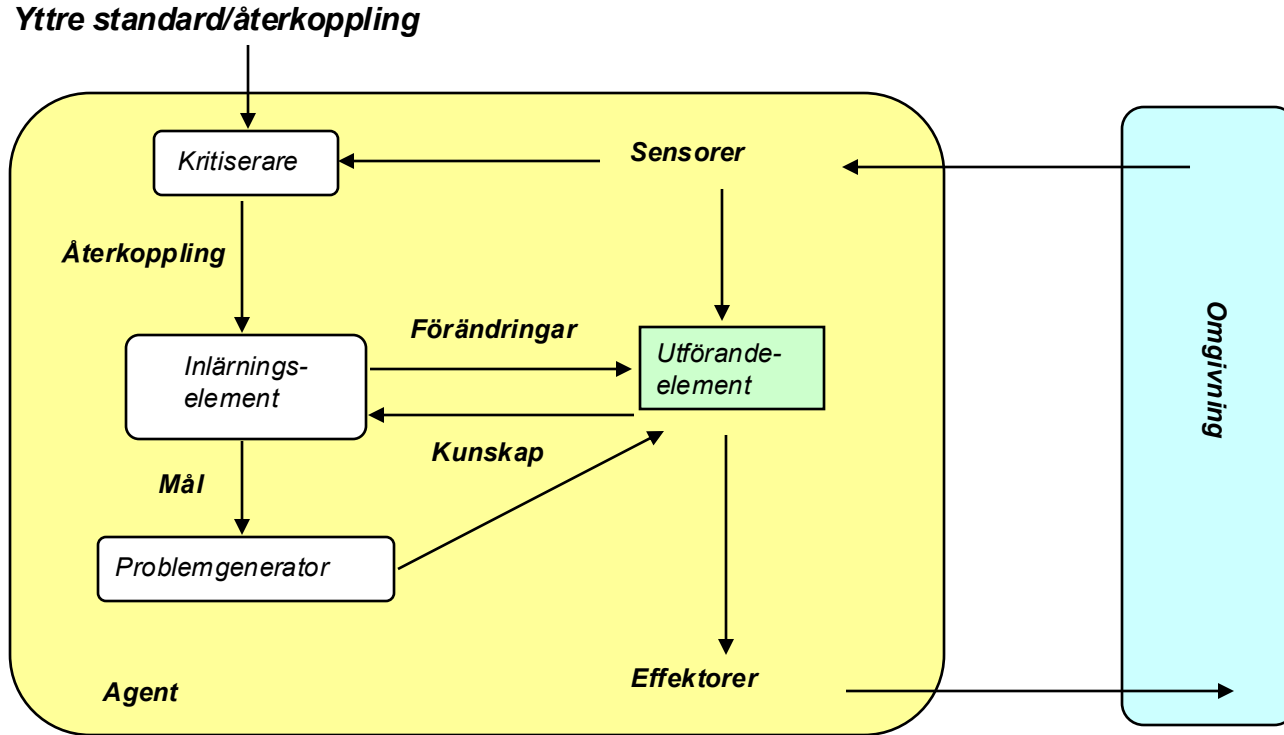
Maskininlärning

- Introduktion
- Beslutsträdsinlärning
 - Entropi
- Linjär regression
- Vektorer
- Klassificering
 - Perceptroner
- Artificiella Neurala Nät (ANN)

Maskininlärning

- Introduktion
- Beslutsträdsinlärning
 - Entropi
- Linjär regression
- Vektorer
- Klassificering
 - Perceptroner
- Artificiella Neurala Nät (ANN)

Agent som lär sig



Typer av träning

- Övervakad inlärning (*supervised learning*)
 - Omgivningen talar direkt om vilken handling som skulle varit korrekt eller det finns ett facit för träningsdatan
- Oövervakad inlärning (*unsupervised learning*)
 - Ingen återkoppling från omgivningen
 - Olika former av klustring eller gruppering av värden
- Förstärkt inlärning (*reinforcement learning*)
 - Agenten får bara reda på om det var bra eller dåligt, inte varför eller vilken handling som orsakade detta

Övervakad inlärning

- Givet en träningsmängd $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ där x är indata och y är facit
- Anta att det finns någon okänd funktion som kan hitta svaret, $f(x) = y$
- Uppgiften är att hitta en funktion h som approximerar f
- h är en hypotes om världen, en **modell** av världen
- y kallas facit, guldstandard (*ground truth*)

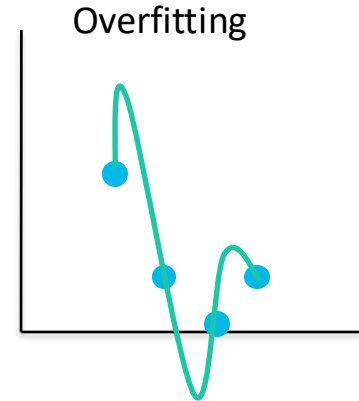
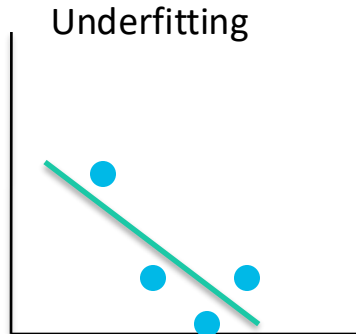
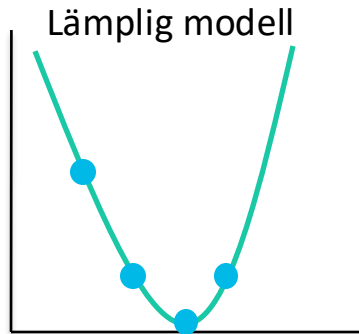
Inläarning

Datamängd med indata, x och motsvarande korrekta målvärde y . En del används till träning, en del används till testning.

- Träning: Visa upp alla x och y och låt systemet lära sig
- Testning: Visa upp x och jämför predicerat värde med y
- Träningsfel: Fel hos modellen vid träning
- Testfel: Hur mycket fel modellen gör på okända data (generaliseringsfel)

Underfitting och overfitting

- Underfitting
 - Modellen räcker inte för att prestera bra på träningsmängden. Den är ännu sämre på testmängden.
- Overfitting
 - Modellen är överoptimerad på träningsdata. Lågt felvärde på träningsdata men högt på testdata.



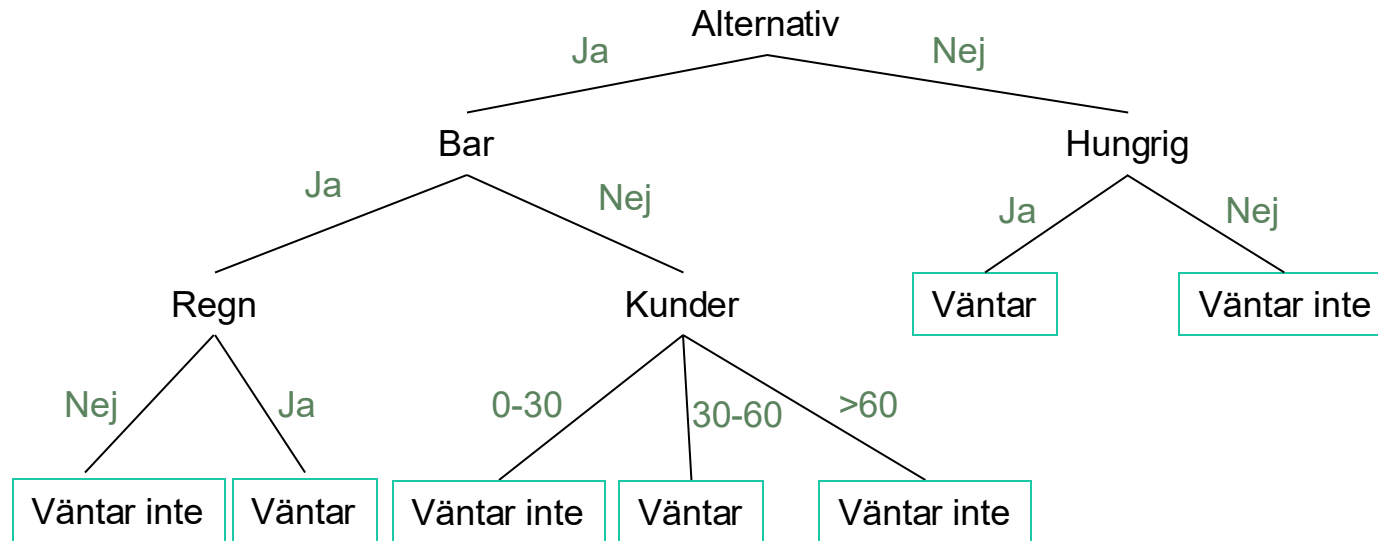
Utvärdering

- Hur bra generaliserar modellen, h
 - Bias: hur mycket avviker modellen och på vilket sätt
 - Varians: hur olika blir modellen beroende på träningsdata
- Avvägning mellan bias och varians
 - Liten bias kan ge komplex modell som passar träningsdata perfekt
 - Liten varians kan ge enklare modeller som generaliserar bättre
- Ofta vill man välja den enklaste modell som passar data

Beslutsträd (*decision trees*)

- Skapa en kompakt representation av beslut i en situation beskriven som en mängd egenskaper som påverkar beslutet
- Ex från kursboken om att vänta på restaurang (Russel & Norvig)
 - Attribut
 - Finns det alternativ i närheten
 - Finns det en bar att vänta i
 - Är det helg
 - Är jag hungrig
 - Finns det kunder i restaurangen
 - Regnar det
 - Vilken typ av restaurang är det
 - Etc.

Restaurangexemplet



Beslutsträdsinläarning

Skapa optimalt beslutsträd ur en exempelmängd

Initialisera rotnoden med hela exempelmängden

while det finns en nod, N, som kan delas:

for each attribut A evaluera delning vid A:

 använd bästa A för att dela N

Exempel: Är personen gift?

NAMN	INKOMST	ÅLDER	GIFT
Marie	55000	28	JA
Sven	40000	31	JA
Olle	30000	37	NEJ
Eva	55000	42	JA
Uwe	35000	27	NEJ
Hans	30000	36	NEJ
Anne	60000	53	JA

Konvertera numeriska data

NAMN	INKOMST	ÅLDER	GIFT
Marie	≥ 40000	< 40	JA
Sven	≥ 40000	< 40	JA
Olle	< 40000	< 40	NEJ
Eva	≥ 40000	≥ 40	JA
Uwe	< 40000	< 40	NEJ
Hans	< 40000	< 40	NEJ
Anne	≥ 40000	≥ 40	JA

Dela vid ålder?

NAMN	INKOMST	ÅLDER	GIFT
Marie	≥ 40000	< 40	JA
Sven	≥ 40000	< 40	JA
Olle	< 40000	< 40	NEJ
Eva	≥ 40000	≥ 40	JA
Uwe	< 40000	< 40	NEJ
Hans	< 40000	< 40	NEJ
Anne	≥ 40000	≥ 40	JA

< 40

NAMN	INKOMST	GIFT
Marie	≥ 40000	JA
Sven	≥ 40000	JA
Olle	< 40000	NEJ
Uwe	< 40000	NEJ
Hans	< 40000	NEJ

≥ 40

NAMN	INKOMST	GIFT
Eva	≥ 40000	JA
Anne	≥ 40000	JA

Måste dela igen!

Dela vid inkomst?

NAMN	INKOMST	ÅLDER	GIFT
Marie	□ 40000	< 40	JA
Sven	□ 40000	< 40	JA
Olle	< 40000	< 40	NEJ
Eva	□ 40000	□ 40	JA
Uwe	< 40000	< 40	NEJ
Hans	< 40000	< 40	NEJ
Anne	□ 40000	□ 40	JA

< 40000

≥ 40000

NAMN	ÅLDER	GIFT
Olle	< 40	NEJ
Uwe	< 40	NEJ
Hans	< 40	NEJ

NAMN	ÅLDER	GIFT
Marie	< 40	JA
Sven	< 40	JA
Eva	≥ 40	JA
Anne	≥ 40	JA

Information

- Vilket attribut delar trädet bäst?
- Ofta används informationsmått
- Information innebär *minskad osäkerhet*
- Exempel
 - Om man i ett varuhus frågar var skivavdelningen finns och får veta att det finns på våning 1, får man mer information om det finns 8 våningar än om det finns 2 men ingen information om det bara finns en våning.
- Utfallet av en händelse med flera lika sannolika möjligheter ger mer information än utfallet av en händelse med färre lika sannolika möjligheter

Logaritmer (recap)

- $10^3 = 1000$
- Vad behöver vi upphöja 10 med för att det ska bli 1000?
 $\log_{10}(1000) = 3$
- $2^{14} = 16\,384$
Vad behöver vi upphöja 2 med för att det ska bli 16 384?
 $\log_2(16\,384) = 14$
- $\log_2(0,63) \approx -0,667$

Entropi

- Hur mycket varians eller variation finns i datan?
- Anta att det finns C olika klasser i vår datamängd d . Entropi kan då beräknas som:

$$E(d) = - \sum_{c \in C} p_c \log_2(p_c)$$

Entropi för datamängden d

Summan för alla klasser

Sannolikheten att ett slumpmässigt element i d tillhör klassen c

$$E(d) = - \sum_{c \in \mathcal{C}} p_c \log_2(p_c)$$

Entropi, exempel 1

Blå eller grön med 50% sannolikhet

$$E(d) = -\left(0,5 * \log_2(0,5) + 0,5 * \log_2(0,5)\right) = -\log_2(0,5) = 1$$

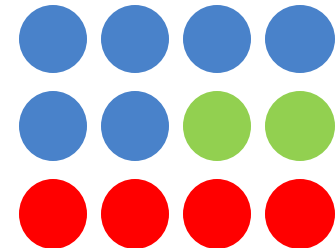
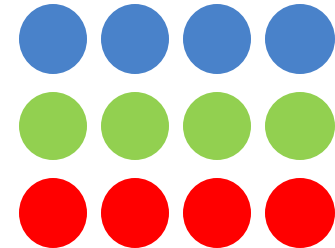
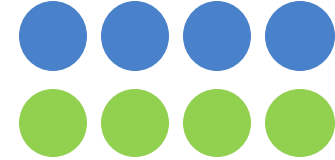
Blå, grön eller röd med en sannolikhet på 1/3

$$E(d) = -\left(\frac{1}{3} \log_2 \frac{1}{3} + \frac{1}{3} \log_2 \frac{1}{3} + \frac{1}{3} \log_2 \frac{1}{3}\right) = -\log_2 \frac{1}{3} \approx 1,584$$

Blå, grön eller röd med 1/2, 1/6 respektive 1/3 sannolikhet

$$E(d) = -\left(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{6} \log_2 \frac{1}{6} + \frac{1}{3} \log_2 \frac{1}{3}\right) \approx 1,459$$

d



Entropi, exempel 2

- Musikavdelningen i varuhus med två lika sannolika våningar

$$-(\frac{1}{2}\log_2(\frac{1}{2}) + \frac{1}{2}\log_2(\frac{1}{2})) = -\log_2(\frac{1}{2}) = \log_2(2) = 1$$

- 8 lika sannolika våningar

$$-\frac{1}{8}\log_2(\frac{1}{8}) - \dots - \frac{1}{8}\log_2(\frac{1}{8}) = -\log_2(\frac{1}{8}) = \log_2(8) = 3$$

- Jämför vanlig datarepresentation

1 bit för att representera 2 olika alternativ, 1 eller 0

3 bit för att representera 8 olika alternativ, 100 eller 010 eller 001

Välja bästa attribut

- Vi vill välja det attribut som ger mest information, dvs. det som *minskar* entropin så mycket som möjligt
- Informationsteori
 - En *bit* information motsvarar ett ja/nej svar på en fråga där båda svaren är lika sannolika
 - Informationsinnehållet beräknas på liknande sätt som entropin men viktat med avseende hur data delas upp av ett attribut

Information för attribut A

Entropin för attributvärden för A beräknas utifrån andelen instanser som hamnar i varje gren och entropin för respektive gren

d_v : instanserna i gren v

d : inkommande instanser

d_{vc} : instanser i gren v av klass c

$$I(A, d) = \sum_{v \in A} \frac{|d_v|}{|d|} \left[- \sum_{c \in C} p_c \log_2 p_c \right]$$

Sannolikheten att ett slumpmässigt element hamnar i grenen v

Entropin i respektive gren

Informationsvinst (*information gain*)

- Beräknas lite olika beroende på vilken algoritm som används
- Men generellt
 - Den information som man vinner om man går ner i grenarna för ett attribut
 - Eller med hur mycket entropin (osäkerheten) minskas
- $Gain(A, d) = E(d) - I(A, d)$
- Välj det attribut som minskar entropin (osäkerheten) så mycket som möjligt

Exempel från tidigare

Mål: gift eller ogift?

NAMN	INKOMST	ÅLDER	GIFT
Marie	≥ 40000	< 40	JA
Sven	≥ 40000	< 40	JA
Olle	< 40000	< 40	NEJ
Eva	≥ 40000	≥ 40	JA
Uwe	< 40000	< 40	NEJ
Hans	< 40000	< 40	NEJ
Anne	≥ 40000	≥ 40	JA

$$E(d) = - \sum_{c \in C} p_c \log_2(p_c)$$

Starttillstånd

- 4/7 är gifta
- 3/7 är ogifta

$$E(d) = - \left(\frac{4}{7} \log_2 \frac{4}{7} + \frac{3}{7} \log_2 \frac{3}{7} \right) = 0,985$$

$$E(d) = - \sum_{c \in C} p_c \log_2(p_c)$$

Ålder

- < 40 (2 gifta, 3 ogifta)
- ≥ 40 (2 gifta, 0 ogifta)

NAMN	INKOMST	ÅLDER	GIFT
Marie	≥ 40000	< 40	JA
Sven	≥ 40000	< 40	JA
Olle	< 40000	< 40	NEJ
Eva	≥ 40000	≥ 40	JA
Uwe	< 40000	< 40	NEJ
Hans	< 40000	< 40	NEJ
Anne	≥ 40000	≥ 40	JA

< 40

≥ 40

NAMN	INKOMST	GIFT
Marie	≥ 40000	JA
Sven	≥ 40000	JA
Olle	< 40000	NEJ
Uwe	< 40000	NEJ
Hans	< 40000	NEJ

NAMN	INKOMST	GIFT
Eva	≥ 40000	JA
Anne	≥ 40000	JA

$$-\left(\frac{2}{2} \log_2 \frac{2}{2}\right) = 0$$

$$-\left(\frac{2}{5} \log_2 \frac{2}{5} + \frac{3}{5} \log_2 \frac{3}{5}\right) = 0,971$$

Beräkna entropi för respektive gren

Ålder, forts.

< 40: 5 av 7 med entropi 0,971

≥ 40: 2 av 7 med entropi 0

$$I(\text{ålder}, d) = \frac{5}{7} \times 0,971 + \frac{2}{7} \times 0 = 0,693$$

$$\text{Gain}(\text{ålder}, d) = 0,985 - 0,693 = 0,292$$

$$I(A, d) = \sum_{v \in A} \frac{|dv|}{|d|} \left[\sum_c - \frac{|d_{vc}|}{|d_v|} \log_2 \left(\frac{|d_{vc}|}{|d_v|} \right) \right]$$

Inkomst

- < 40000 (0 gifter, 3 ogifter)
- ≥ 40000 (4 gifter, 0 ogifter)

$$I(\text{inkomst}, d) = \frac{3}{7} \times 0 + \frac{4}{7} \times 0 = 0$$

$$\text{Gain}(\text{ålder}, d) = 0,985 - 0 = 0,985$$

NAMN	INKOMST	ÅLDER	GIFT
Marie	≥ 40000	< 40	JA
Sven	≥ 40000	< 40	JA
Olle	< 40000	< 40	NEJ
Eva	≥ 40000	≥ 40	JA
Uwe	< 40000	< 40	NEJ
Hans	< 40000	< 40	NEJ
Anne	≥ 40000	≥ 40	JA

< 40000

≥ 40000

NAMN	ÅLDER	GIFT
Olle	< 40	NEJ
Uwe	< 40	NEJ
Hans	< 40	NEJ

$$\left(\frac{3}{3} \log_2 \frac{3}{3}\right) = 0$$

NAMN	ÅLDER	GIFT
Marie	< 40	JA
Sven	< 40	JA
Eva	≥ 40	JA
Anne	≥ 40	JA

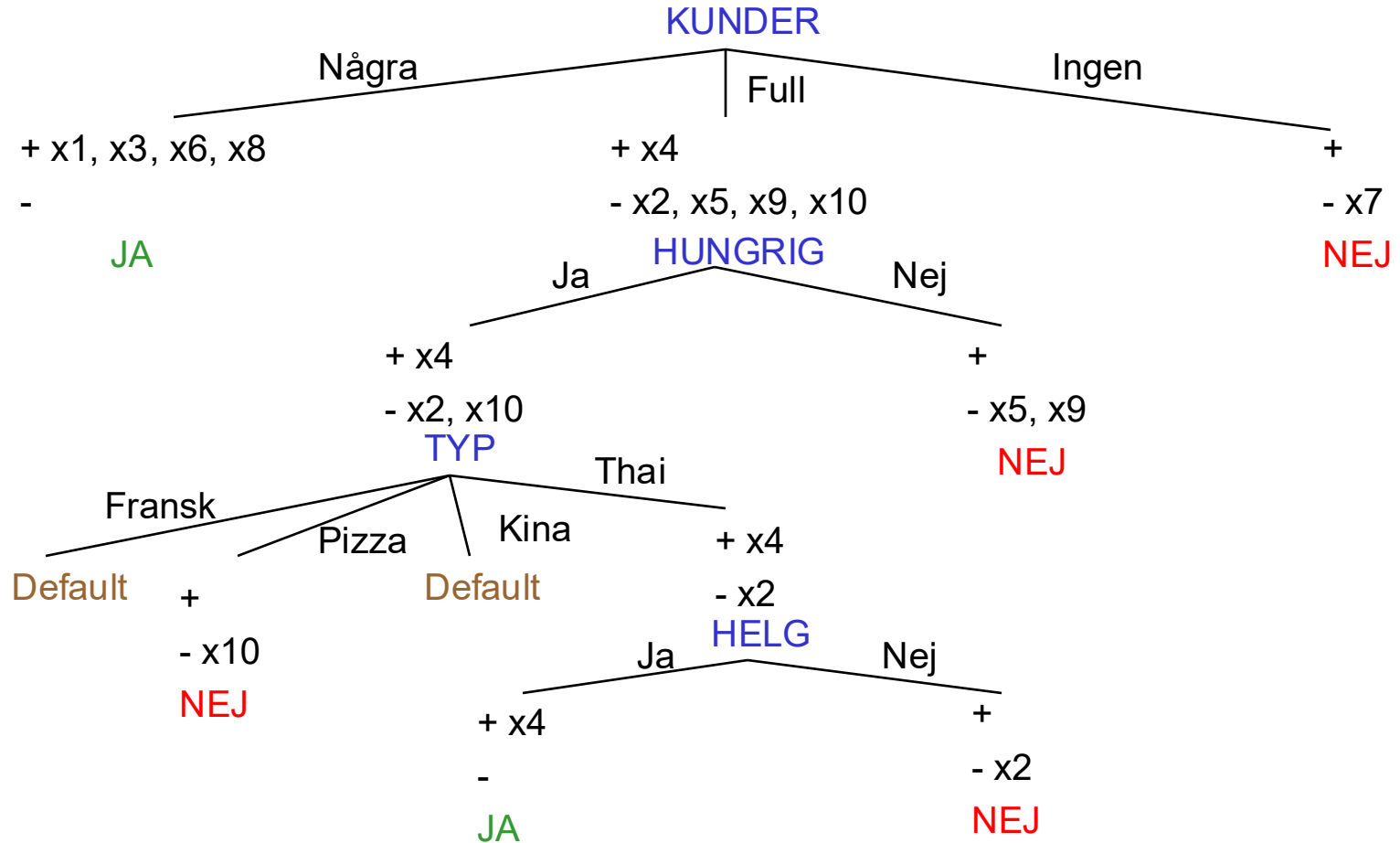
$$\left(\frac{4}{4} \log_2 \frac{4}{4}\right) = 0$$

Restuarangexemplet, träningsexempel

Exempel	Attribut									Mål Väntar
	Alt	Bar	Helg	Hunger	Kunder	V-tid	Regn	Bokat	Typ	
X1	Ja	Nej	Nej	Ja	Några	0-10	Nej	Ja	Fransk	Ja
X2	Ja	Nej	Nej	Ja	Full	30-60	Nej	Nej	Thai	Nej
X3	Nej	Ja	Nej	Nej	Några	0-10	Nej	Nej	Kina	Ja
X4	Ja	Nej	Ja	Ja	Full	10-30	Nej	Nej	Thai	Ja
X5	Ja	Nej	Ja	Nej	Full	>60	Nej	Ja	Fransk	Nej
X6	Nej	Ja	Nej	Ja	Några	0-10	Ja	Ja	Pizza	Ja
X7	Nej	Ja	Nej	Nej	Ingen	0-10	Ja	Nej	Kina	Nej
X8	Nej	Nej	Nej	Ja	Några	0-10	Ja	Ja	Thai	Ja
X9	Nej	Ja	Ja	Nej	Full	>60	Ja	Nej	Kina	Nej
X10	Ja	Ja	Ja	Ja	Full	10-30	Nej	Ja	Pizza	Nej

+: väntar
-: väntar inte

Beslutsträd



Resultat

- Beslutsträdet använder inte alla attribut
- Kan ha funnit okända samband mellan attribut
- Om det saknas exempel för ett visst attributvärde så returneras något defaultvärde
- Om attributen inte räcker för att det ska finnas beslut i alla grenar i träningsdatan så saknas det antingen information och fler attribut behövs eller så är det ett fel

Sammanfattning

- Olika typer av återkoppling i maskininlärning
 - Övervakad inlärning
 - Oövervakad inlärning
 - Förstärkt inlärning
- Beslutsträdsinlärning
- Informationsbegrepp
 - Högre entropi, mer osäkerhet
 - Lägre entropi, mindre osäkerhet
- Beslutsträd byggs upp genom att välja de attribut som ger mest information